Cassandra Crossing/ Archivismi: API, quando il gioco si fa duro

(563)—Oggi ci sposteremo su un differente piano di utilizzo di Internet Archive, quello della "programmazione" via API; ma prima dovremo…

Cassandra Crossing/ Archivismi: API, quando il gioco si fa duro
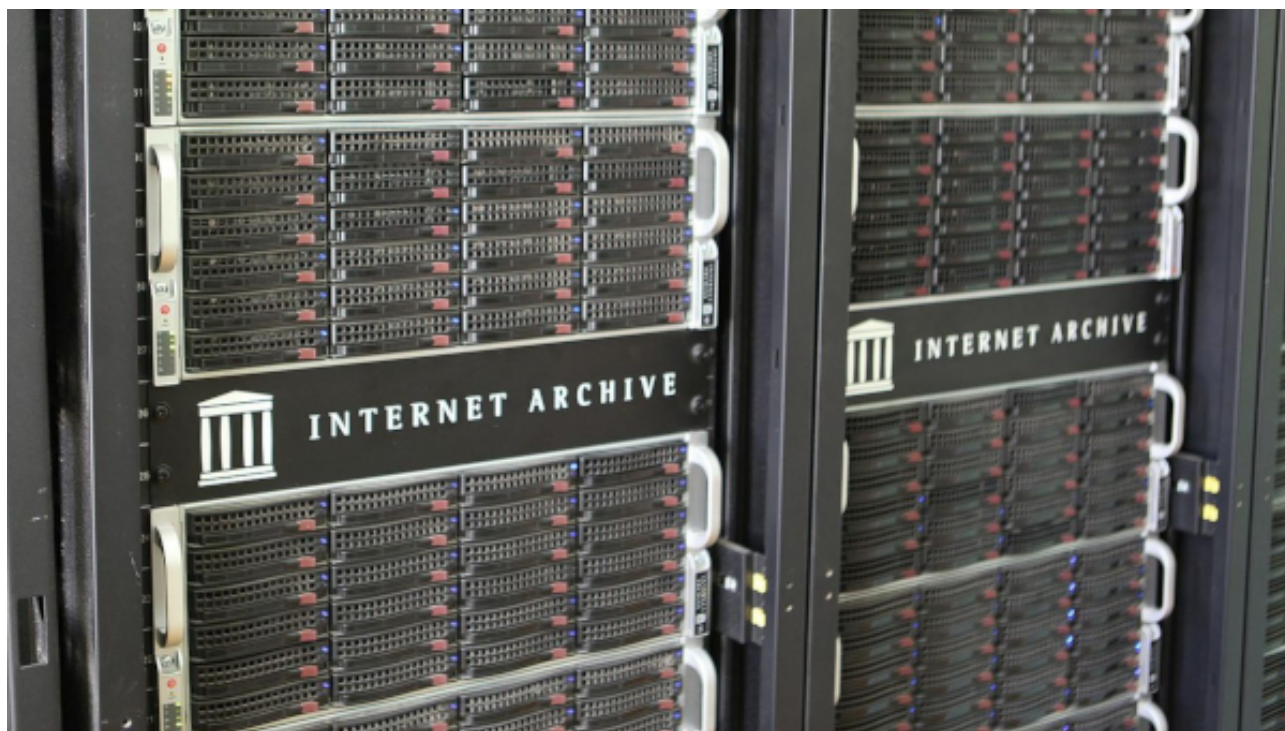
Figure 1:

(563)—Oggi ci sposteremo su un differente piano di utilizzo di Internet Archive, quello della "programmazione" via API; ma prima dovremo obbligatoriamente parlare di doveri e responsabilità degli utenti di Internet Archive.

29 dicembre 2023—Nelle ultime due puntate (è da oggi disponibile una lista completa degli articoli di "Archivismi") ci siamo occupati di un'archiviazione elementare su *Internet Archive*; archiviare un singolo file ci ha comunque aperto una parte significativa del sistema che abbiamo davanti, e delle potenti funzionalità che ci mette a disposizione.

Molto, molto altro rimane da mostrare, anche solo per le operazioni di archiviazione manuali. Prossimamente descriveremo e realizzeremo una vera campagna di archiviazioni, raccontando le minuzie ed i problemi spiccioli che distinguono i casi reali dagli esempi che troviamo sui manuali.

Ma oggi tratteremo di un argomento già accennato di sfuggita in una puntata precedente, e che porta la potenza archivistica che *Internet Archive* mette a disposizione dei suoi utenti ad un nuovo livello. Parliamo ovviamente della possibilità di "programmare" le operazioni su *Internet Archive*.

Non ci vuole un genio per immaginare che un servizio come *Internet Archive* esiste perché ha alle spalle un piccolo esercito di programmatori che scrivono, mantengono e fanno evolvere una base di software dedicato. E per inciso, per fomentare la mai estinta "*Classifica dei migliori linguaggi di programmazione*", anche in *Internet Archive* *Python* la fa da padrone!

Ma torniamo all'argomento di oggi.

In breve: sì, è possibile usare Internet Archive usando script o veri programmi che automatizzano le operazioni di archiviazione che decidiamo di realizzare.

E sì, questo viene realizzato "*esponendo una API*". Per il conforto dei non programmatori, significa semplicemente che è possibile automatizzare le operazioni da compiere utilizzando degli script o dei veri e propri programmi, che eseguono, ovviamente via Rete, precise chiamate a delle funzionalità elementari di Internet Archive, definite appunto in una API—*Interfaccia per la Programmazione Applicativa*.

Non ci sarebbe bisogno di dire altro, semplicemente di fornire nuovamente il link al *Portale degli Sviluppatori di Internet Archive*, e lasciare che chiunque abbia mai *trafficato*, anche solo realizzando uno script .bat per il DOS, scopra ed utilizzi la potenza delle API di Internet Archive.

Ma no, un minimo di indicazioni e raccomandazioni preliminari sono comunque necessarie, prima di fare anche solo un piccolissimo esempio.

In primis, **Internet Archive non pone limiti predefiniti a quello che un utente può fare dei servizi che vengono forniti**; ad esempio non limita a priori la quantità di informazioni che possono essere archiviate.

Ma nessuna realtà esposta al pubblico può essere "*indifesa*", visto che una percentuale di imbecilli, profittatori e delinquenti esistenti al mondo è presente anche tra gli utenti di Internet Archive.

Come la storia della Rete ha più volte dimostrato, realtà collaborative di grandi dimensioni, ad esempio Wikipedia, riescono a sopravvivere e svilupparsi solo se gestite come un ibrido tra democrazia imperfetta e tirannia illuminata. *Internet Archive* non fa eccezione.

E' per questo che alcune risorse, come ad esempio le *Collezioni*, vengono centellinate e fornite solo a richiesta. Una serie di amministratori di vario livello supervede e controlla infatti il funzionamento e l'utilizzo di Internet Archive, e tiene in riga, bacchetta od espelle gli utenti *disfunzionali*. **Una tale presenza non deve essere vista come un problema od un limite, ma come una risorsa**; infatti gli amministratori hanno il ruolo principale di aiutare tutti gli utenti ad utilizzare *Internet Archive*.

Gli amministratori sono tuttavia una risorsa preziosa e scarsa; **mandare una email agli amministratori**, quando non direttamente previsto dalle procedure (ad esempio per la creazione di una Collection) **deve essere vista come ultima risorsa**, da utilizzare solo dopo un'attenta lettura della documentazione e dell'help in linea, molte prove, una ricerca nel blog e perché no, anche sui normali motori di ricerca. **Mi raccomando!**

Ma non si era detto che avremmo *programmato* qualcosa? Verissimo, e passiamo subito alla pratica. E per partire da qualcosa di semplice ed innocuo, ipotizziamo di aver trovato una serie di cose che ci interessano, ad esempio parecchi numeri di una rivista, e di volerli scaricare in maniera veloce, affidabile, e che non richieda operazioni manuali ripetitive.

E per semplicità, faremo il tutto da linea comandi, senza utilizzare direttamente le API e quindi senza dover scrivere un vero programma in Python o simili; ci basterà scaricare il programma

Python “***ia***” ed utilizzarlo. ***ia*** è un programma già”pseudo-compilato”, cioè scritto in un “linguaggio” intermedio detto *Python Bytecode*, che è portabile su qualsiasi piattaforma abbia un ambiente Python3 installato.

L’utilizzo di una versione di Linux, Debian, Ubuntu etc., è vivamente consigliato. Potete anche utilizzarlo in una macchina virtuale Virtualbox o VMWare su qualsiasi computer.

Dovrebbe anche funzionare l’ambiente WSL di Windows, ma qui Cassandra non procede oltre ed abbandona gli arditi che volessero cimentarsi; anzi, eventualmente aspetta da loro dei feedback a riguardo per integrare questo articolo.

Quindi torniamo con Cassandra alla sua amata Debian, ed installiamo e configuriamo *ia* con la procedura che troviamo qui. Ma anche un semplice

> sudo apt install internetarchive

è sufficiente. Miracoli di Debian …

In breve, su un computer dove sia installato l’ambiente Python3 si deve scaricare dove preferiamo, oppure installare, il comando ia, renderlo eseguibile, ed infine lanciarlo con il parametro *configure* per associarlo al nostro utente (avete creato il vostro utente, vero?).

E’ tutto pronto; come primo esempio con il seguente comando possiamo scaricare il solo pdf originale del nostro articolo di esempio, che avevamo caricato la scorsa puntata.

> $ ./ia download cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi—no-directories—format=“Text PDF”

> cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi:
> downloading Cassandra_Crossing_2558_Il Dizionario di Cassandra_ Archivismi.pdf:
> 100%| | 513k/513k [00:00<00:00, 709kiB/s

Ma se avessimo voluto scaricare tutto l’oggetto, file derivativi inclusi, avremmo potuto scrivere ancor più semplicemente

> $ ./ia download cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi

Avremmo così ottenuto una directory con lo stesso nome dell’identificatore dell’oggetto, contenente tutti i file da cui è formato. Lo stesso procedimento funziona anche per scaricare una intera collezione, o parti di essa. Un’altra raccomandazione, **calcolate prima quanto è grande la selezione che avete fatto**; su *Internet Archive* ci sono oggetti di dimensioni enormi.

Per avere aiuto, oltre che consultare la guida online, basta dare i comandi

> $ ./ia help

> $ ./ia help download

> $ ./ia help upload

Terminiamo con altre raccomandazioni in ordine sparso.

Se caricate nuovi oggetti, è meglio usare il metodo con foglio elettronico in formato CSV, di cui trovate un esempio qui o nella guida. In questo modo avrete sempre sotto controllo tutti i parametri insieme. Dare tutti i parametri da linea comandi può essere complesso e si possono facilmente commettere errori.

Quando creerete i vostri oggetti, **inseriteli sempre nella collezione *test_collection***, come è mostrato anche nel foglio di esempio. I motivi li abbiamo già spiegati.

Quando invece inserirete i vostri primi oggetti *definitivi*, non inserite tra i parametri la collection, lasciando quella di default *opensource*. Buona sperimentazione!

Ed anche per oggi è tutto. *Stay tuned* per la prossima puntata di "*Archivismi*".

Scrivere a Cassandra—Twitter—Mastodon
Videorubrica "Quattro chiacchiere con Cassandra"
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

By Marco A. L. Calamari on January 1, 2024.

Canonical link